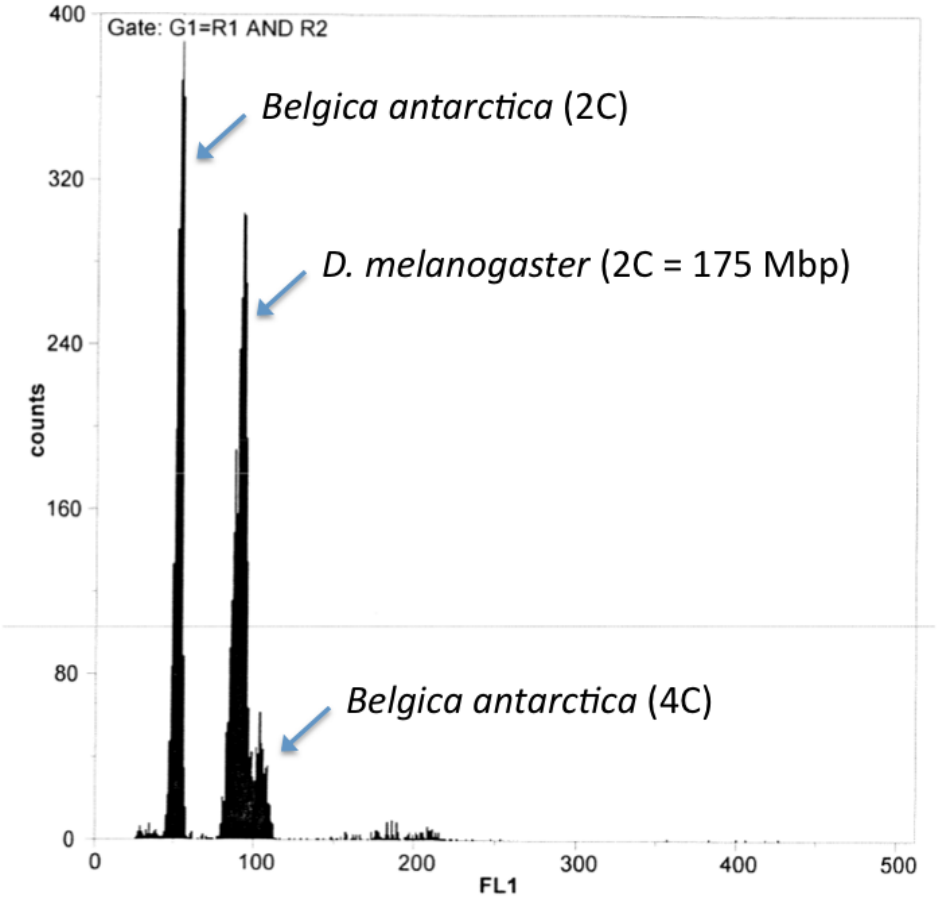
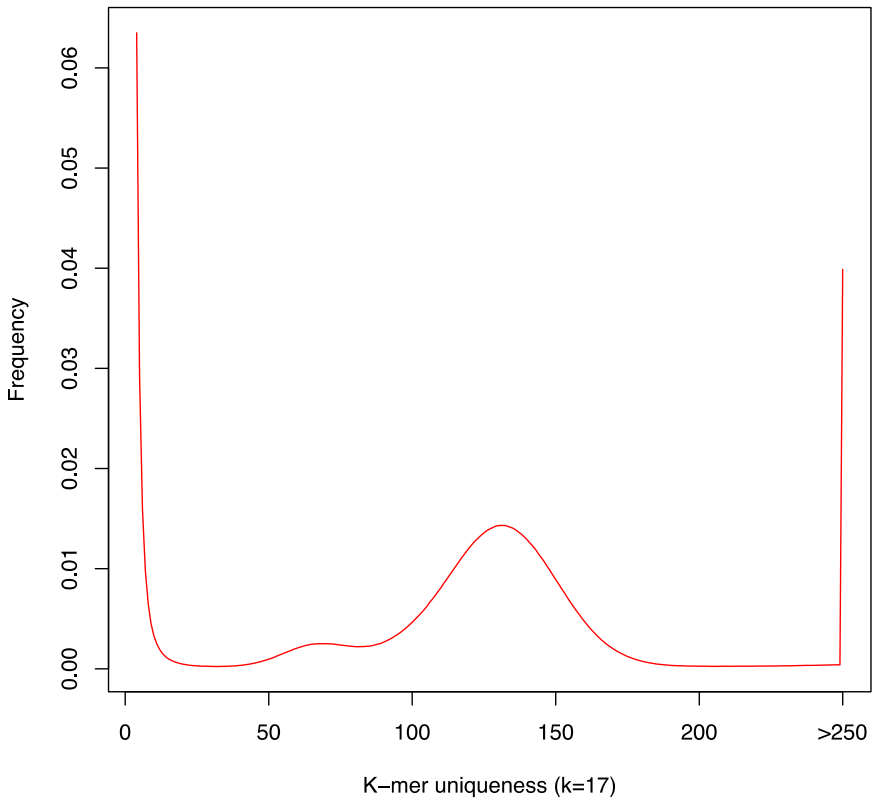
Supplementary Information

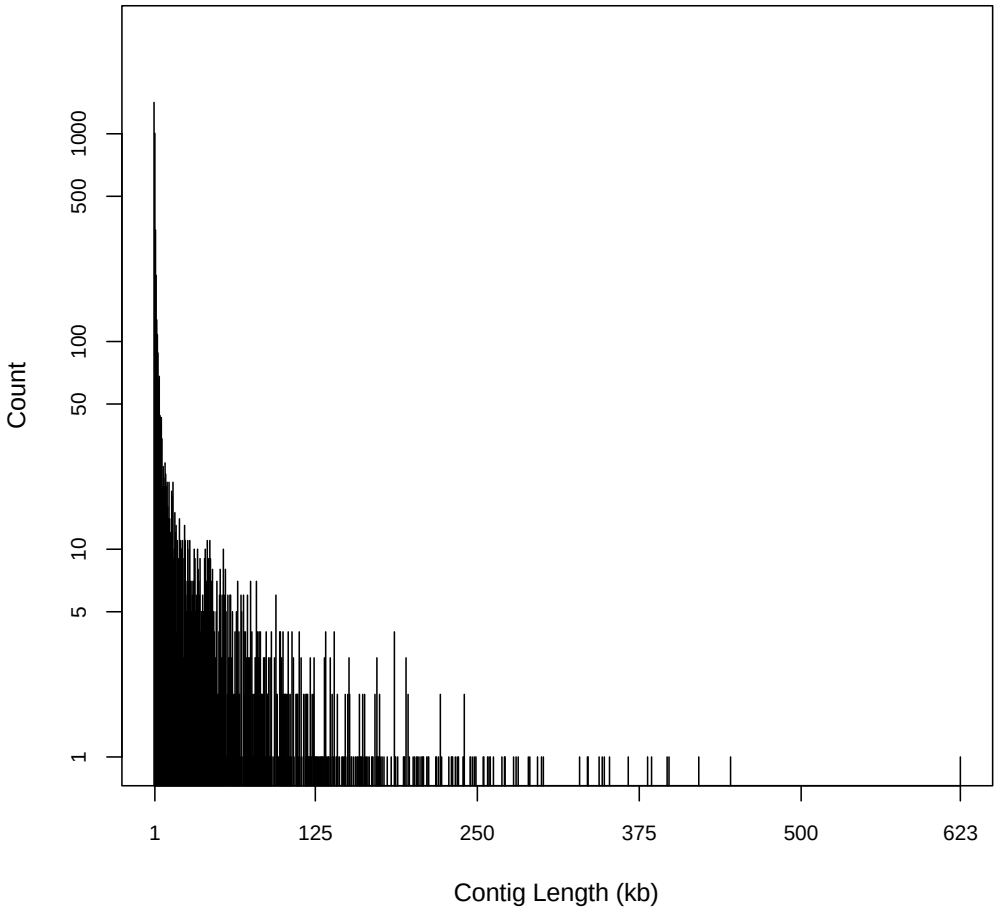
Supplementary Figures



Supplementary Figure 1: Flow cytometry estimates for *B. antarctica.* Flow cytometry estimate histogram for *B. antarctica* using standard of *D. melanogaster*.

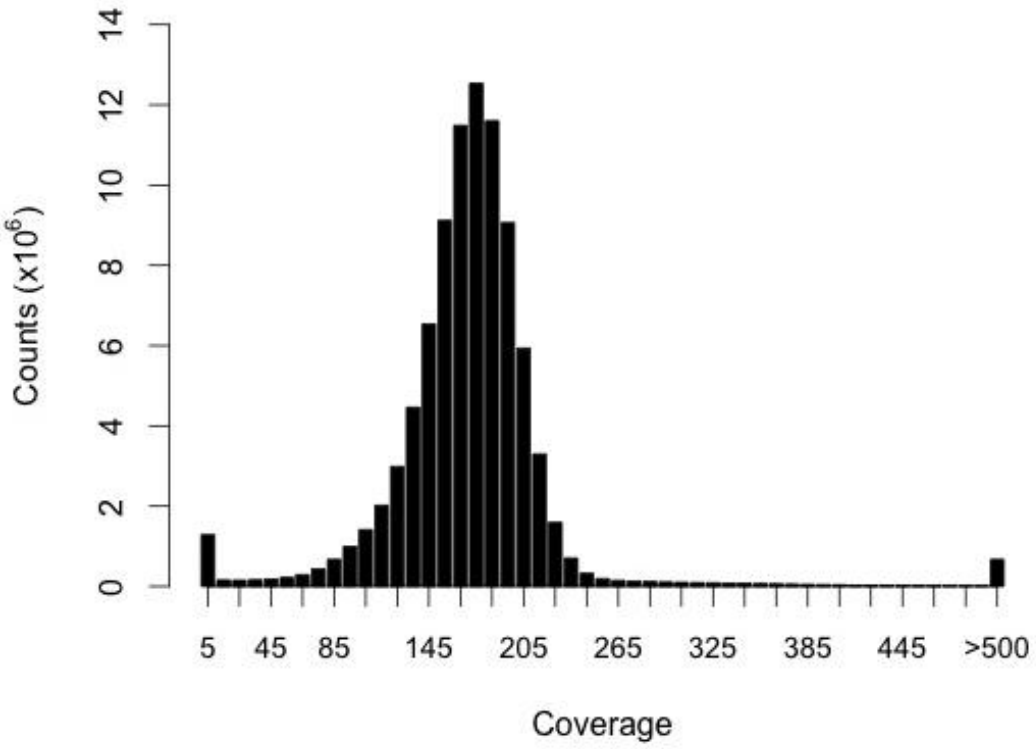


Supplementary Figure 2: Distribution of 17-mers from raw sequence data. Shown here are 17-mers with at least 4 occurrences. For visualization, k-mers that are seen less than 4 times in the dataset are excluded from this plot. Those k-mers likely arise due to sequencing errors. The smaller peak (at uniqueness 69) represents regions in the genome with a single nucleotide polymorphism (SNP).

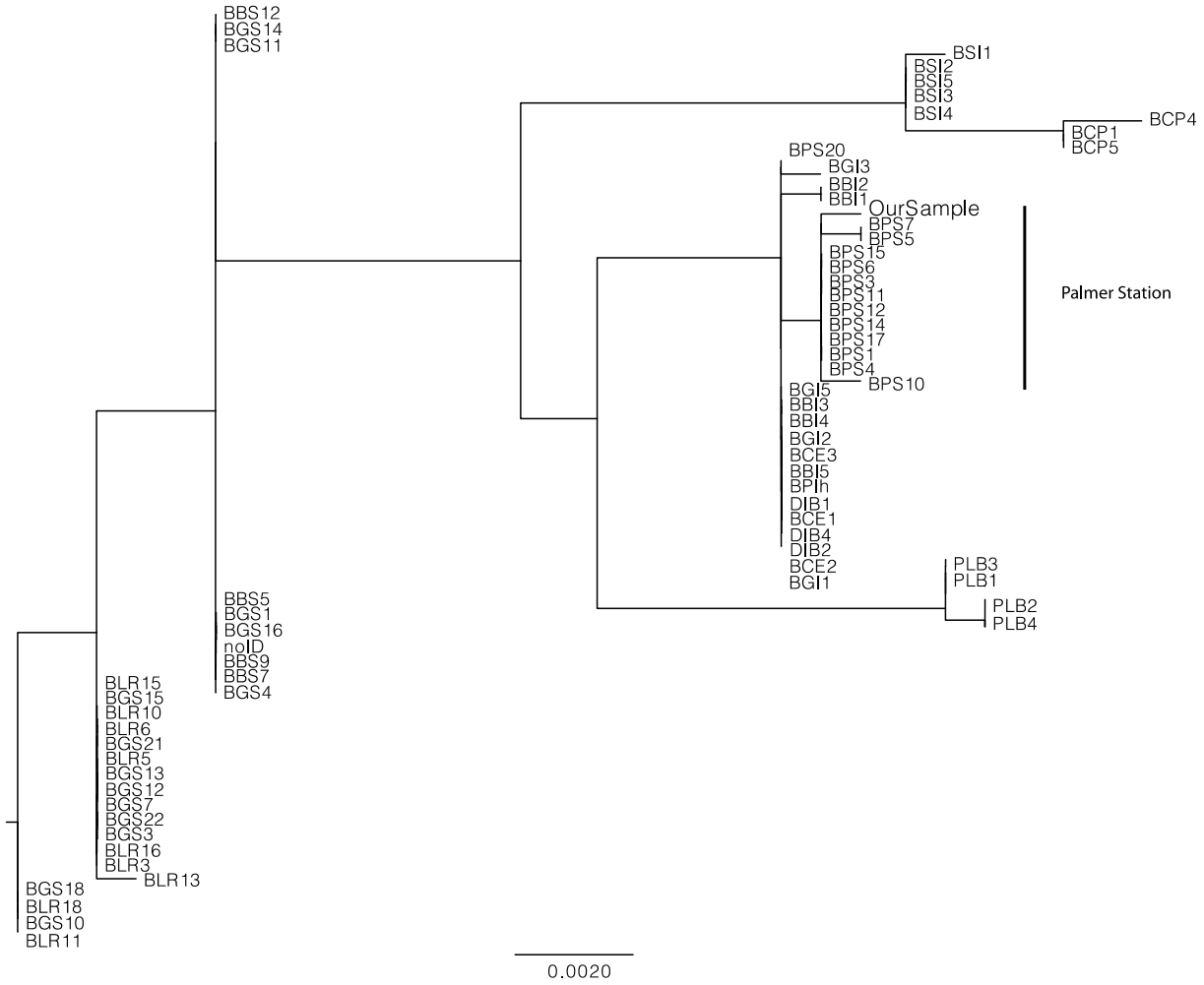


Supplementary Figure 3: **Distribution of scaffold lengths from assembled genome.**

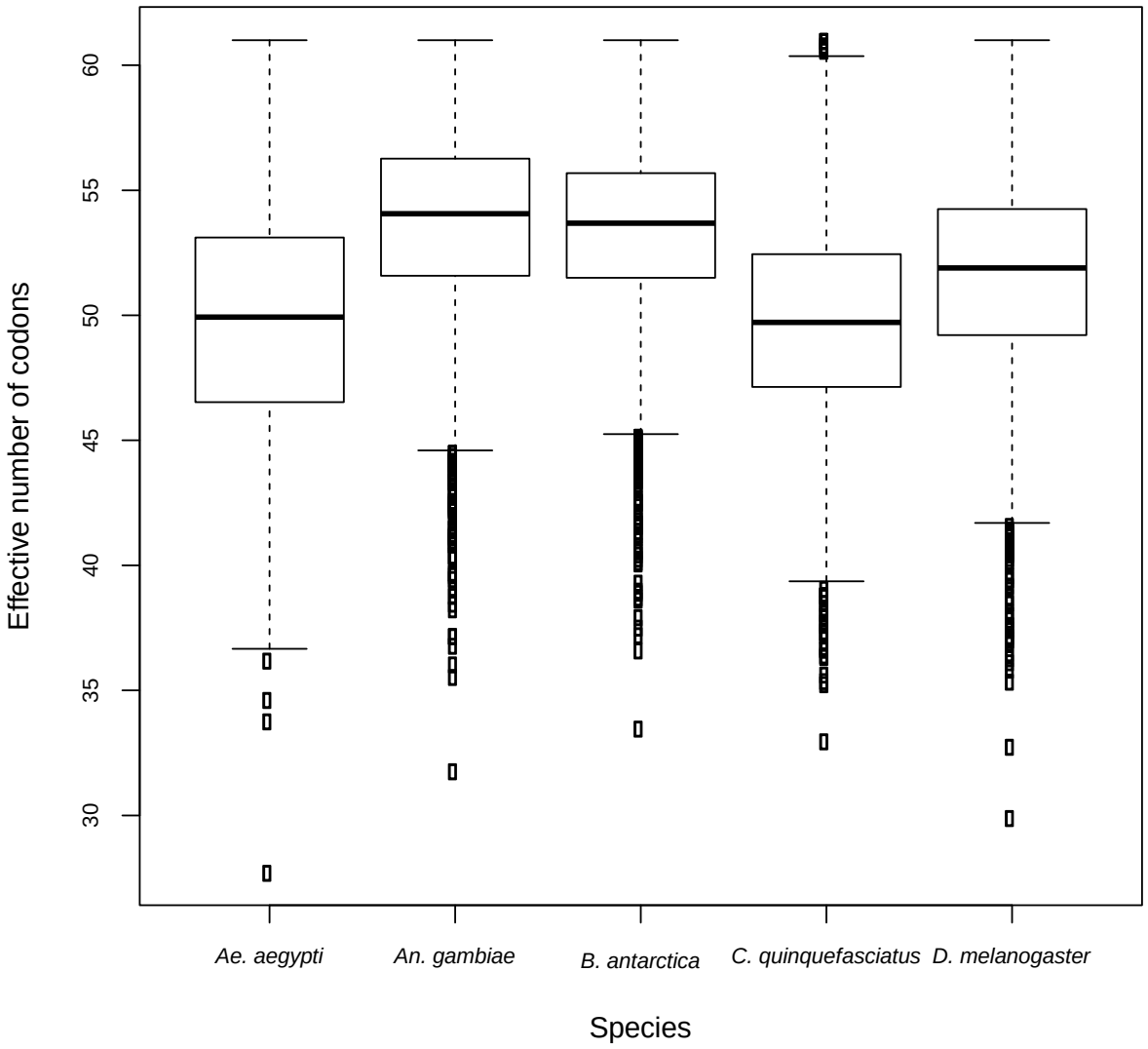
Length distribution of scaffolds from the final genome assembly. The smallest scaffold is 300 base pairs. Note the y-axis is on a log scale.



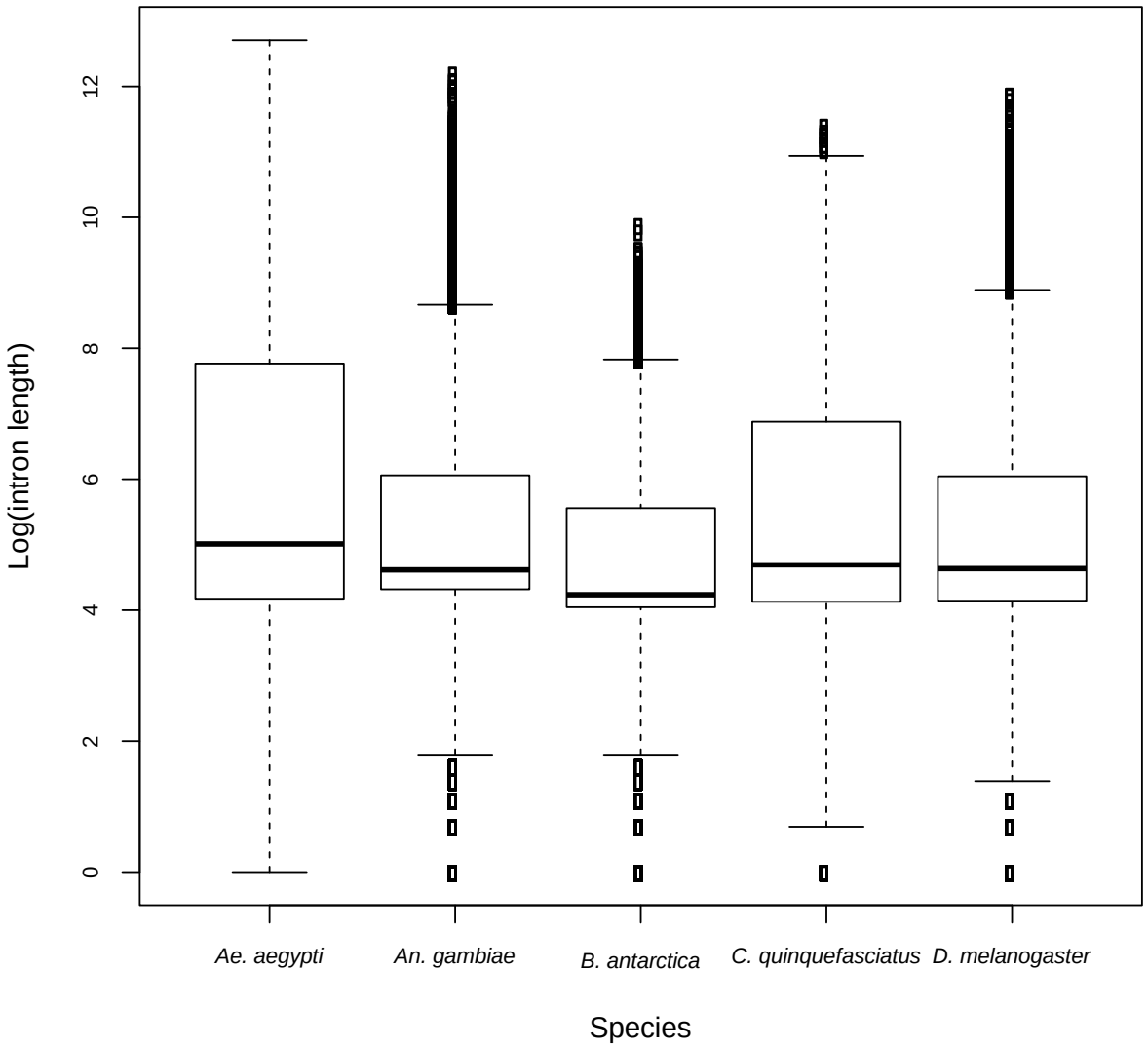
Supplementary Figure 4: Coverage histogram of bases in assembled genome. For each base pair in the assembled genome, coverage is calculated based on reads mapped to the assembled genome using BamTools [1].



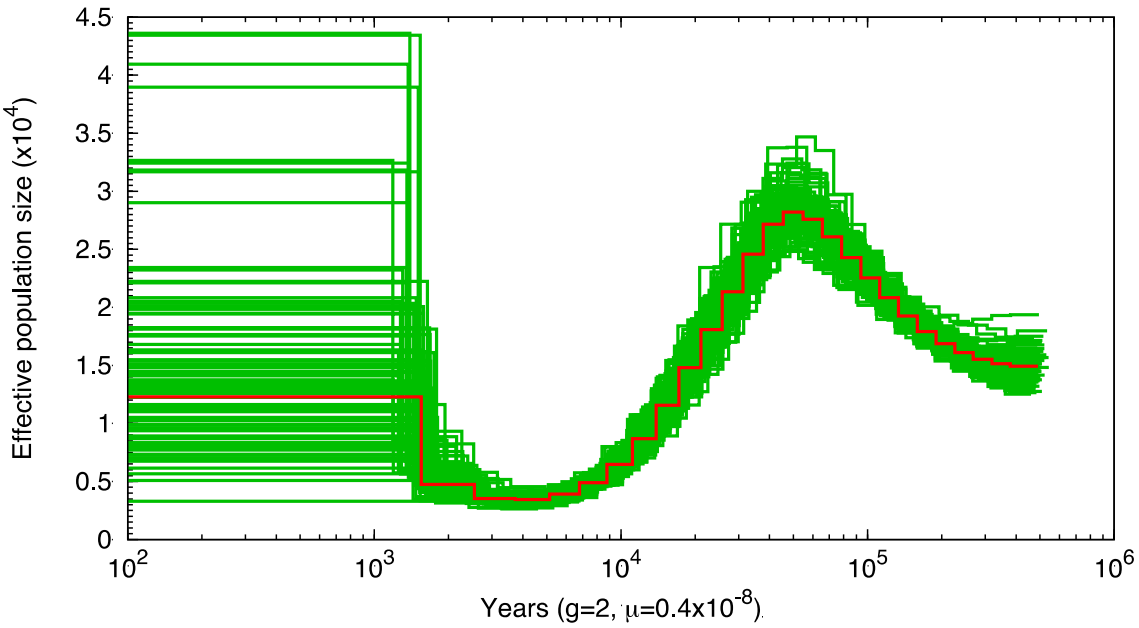
Supplementary Figure 5: *cox1* relationship among *B. antarctica* individuals. Comparison to existing *B. antarctica* sequences from the *cox1* locus collected in the Allegrucci et al. (2012) [2] study was downloaded from Genbank, aligned to the data from this study and a maximum likelihood tree was built using RaxML [3]. Sample codes from Allegrucci et al. (2012) [2] are Livingston Island, Byers Peninsula (BGS, BBS, BLR), Spert Island (BSI), Cierva Point (BCP), Danco Island (DIB), Goudier Island (PLB), Palmer Station (BPS), Berthelot Island (BBI), Peterman Island (BPI), Gand Island (BGI), and Cape Evensen (BCE).



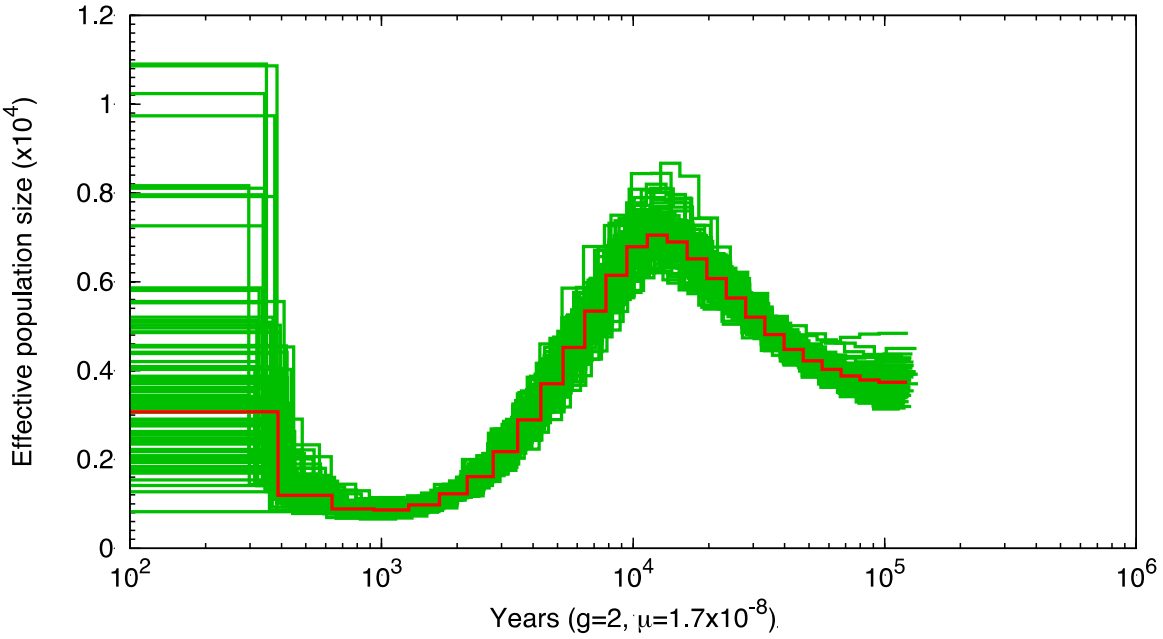
Supplementary Figure 6: Codon usage bias estimates for each species. Side-by-side comparison of boxplots of codon usage among five Diptera species using an estimate of effective number of codons that accounts for background nucleotide composition [4]. Only genes in the set of 3,582 one-to-one orthologs were used in the analysis to ensure that the same loci were compared between species. Each box represents the interquartile range and outliers that are more than or less than 1.5 times the interquartile range are represented as dots in the boxplots.



Supplementary Figure 7: Intron size distribution comparison. Boxplot comparing the natural logarithm of intron size for the five Diptera species. Data also shown in Supplementary Table 12. Each box represents the interquartile range and outliers that are more than or less than 1.5 times the interquartile range are represented as dots in the boxplots.



Supplementary Figure 8: Demographic history inferred from a single *B. antarctica* genome. Pairwise Sequentially Markovian Coalescent (PSMC) analysis for inferred historical population sizes using variant data from the sequenced individual using a mutation rate of 0.4×10^{-8}. The x-axis gives time measured by pairwise sequence divergence converted to years and the y-axis gives the effective population size measured by the scaled mutation rate. The green lines correspond to PSMC inferences on 100 rounds of bootstrapped sequences (as shown in Figure 4), while the red line corresponds to the estimate from the data.



Supplementary Figure 9: Demographic history inferred from a single *B. antarctica* genome. Pairwise Sequentially Markovian Coalescent (PSMC) analysis for inferred historical population sizes using variant data from the sequenced individual using a mutation rate of 1.7×10^{-8}. The x-axis is the time measured by pairwise sequence divergence converted to years and the y-axis gives the effective population size measured by the scaled mutation rate. The green lines correspond to PSMC inferences on 100 rounds of bootstrapped sequences, while the red line corresponds to the estimate from the data.

Supplementary Tables

Supplementary Table 1: Flow cytometry estimates for Chironomidae species

Species	n	Genome size ± S.E. (Mbp)	Reference
Chironomus tentans	-	205	[5]
Chironomus riparius (female)	-	196.2 ± 1.0	[6]
Chironomus riparius (male)	-	194.3 ± 1.1	[6]
Prodiamesa olivacea	-	127	[7]
Allacapnia sp.	3	118.3 ± 1.4	New in this study
Diamesa mendotae	9	116.3 ± 2.2	New in this study
Micropsectra sp.	6	108.2 ± 1.1	New in this study
Belgica antartica (female)	10	99.25 ± 0.4	New in this study
Belgica antarctica (male)	10	98.4 ± 0.1	New in this study

* conversions for published rates based on 1 pg = 978 Mbp

Supplementary Table 2: Comparison of different assemblers.

assembler	version	k	proportion of reads assembled	total assembled bases	number of contigs > 300	N50	NG50	REAPR error free bases (%)	REAPR errors	REAPR warnings
SGA	0.10.13		1/3 data	90448289	15248	33724	29620	82.49	2303	27020
SGA	0.10.13		all	91442312	25571	14861	12876	77.21	2074	42092
ABySS	1.3.7	35	all	88718974	5256	44041	37719	89.90	2192	15016
ABySS	1.3.7	39	all	90460926	5580	41622	37036	89.85	2379	15240
Velvet	1.2.10	49	all	89431929	4870	95812	82919	84.06	9053	18484
velvet	1.2.10	53	all	89969631	5043	95610	82768	83.97	8929	19485
velvet	1.1.05	55	all	89519377	5422	94529	82407	85.88	6203	19889
velvet	1.2.10	55	all	90252044	5089	94510	83802	83.21	9263	20247
velvet	1.2.10	57	all	90660484	5199	95728	85026	82.53	10755	20809
velvet	1.2.10	57	1/3 data	89581777	5434	72821	64577	82.44	10185	20162
velvet+ ERANGE	1.1.05	55	all	89501225	5064	97740	84969	85.85	6427	19956
velvet+ ERANGE+ PacBio	1.1.05		all	89589133	5003	98263	85160	85.93	6440	20464

The quality of different assemblers was compared using REAPR [8], which uses mapping information, including fragment coverage and insert size distribution, to identify putative mis-assemblies. The assemblers compared included SGA [9], ABySS [10] and velvet [11] *de novo* assemblers. For proportion of reads assembled, the data was randomly subset to one-third of the original data to investigate whether the high-coverage was saturating the assembler capabilities. NG50 is similar to N50, however the genome size of 99.25Mb is used for the estimate.

Supplementary Table 3: Nested TEs.

TE orders	TE families	copies (#)
DNA	Mariner2;Tc1	1
DNA	Marwolen1;Mariner2;Tc1	1
DNA	Paris;Quetzal	1
DNA	Quetzal;Tc1	1
DNA	Helitron1;RtaG4	2
DNA;LTR	Galileo;Gypsy68	1
DNA;LTR	P4;HMS-Beagle	1
LTR	Accord2;Gypsy59	1
LTR	Accord2;Stalker	1
LTR	Bel2;ninja	3
LTR	Bel4;diver	2
LTR	Bel4;roo	1
LTR	Bel8;diver	1
LTR	Bel8;diver;max-element	1
LTR	Bel8;max-element	3
LTR	Bel8;rooA	3
LTR	Copia;frogger	1
LTR	Copia;mtanga	5
LTR	Copia1;Copia	2
LTR	Copia1;Copia2	1
LTR	Copia1;Copia4	1
LTR	Copia1;mtanga	3
LTR	Copia2;1731	1
LTR	Copia2;Copia4	2
LTR	Copia4;Copia	5
LTR	Copia4;Copia;frogger	1
LTR	Copia4;frogger	2
LTR	Copia2;frogger	1
LTR	frogger;mtanga	1
LTR	Gypsy10;Gypsy6	1
LTR	Gypsy10;Tabor	1
LTR;non-LTR	Ag-Jock-13;Ag-Outcast-6;mdg1;G2;Tabor	1
LTR;non-LTR	Amer3;Copia	3
LTR;non-LTR	MinoAg1;roo	1
LTR;non-LTR	roo;Rt2	1
LTR;non-LTR	Rt1;Tabor	1
LTR;non-LTR	Tabor;Ag-Jock-13	1
non-LTR	Cr1-1;Cr1-4	1
non-LTR	MinoAg1;Tart	1
non-LTR	R7AG2;Tart	1
non-LTR	Rt1;Tart	1
non-LTR	Rt2;Tart	1
non-LTR	RtaG4;Tart	3

Supplementary Table 4: Distribution of unique TEs insertions within TE orders and families

TE order	TE family	copies (#)
DNA	BuT2	1
	BuT3	1
	DNAREP1	1
	Harbinger1	2
	hAT-2	1
	Helitron1	12
	Helitron2	3
	ISBu2	3
	Kepler	1
	Mariner1	3
	Mariner2	1
	mini-me	3
	P4	1
	Polinton	5
	Quetzal	1
	S-element	1
	Tc1	3
	Transib1	6
	Transib2	1
	Transib3	2
	tsessbeII	1
	Uhu	2
LTR	17.6	3
	1731	1
	Accord	1
	Accord2	2
	aurora-element	1
	Bel1	1
	Bel10	1
	Bel11	2
	Bel12	3
	Bel13	2
	Bel14	5
	Bel15	4
	Bel17	1
	Bel18	2
	Bel2	4
	Bel4	6
	Bel8	3
	Bel9	2
	Burdock	1
	Chouto	1
	Circe	2

TE order	TE family	copies (#)
LTR (continued)	Gypsy55	7
	Gypsy57	3
	Gypsy58	1
	Gypsy59	1
	Gypsy6	3
	Gypsy61	5
	Gypsy62	3
	Gypsy63	10
	Gypsy64	7
	Gypsy65	1
	Gypsy68	23
	Gypsy69	12
	Gypsy8	1
	Gypsy9	1
	HMS-Beagle	5
	Invader1	2
	Invader3	1
	Invader3	1
	max-element	6
	mdg1	6
	mtanga	5
	Ninja	3
	nomad	2
	Osvaldo	4
	roo	8
	rooA	7
	Springer	1
	Stalker2	1
	Tabor	16
	Tom	5
	tv1	1
	Ulysses	1
	ZAM	1
non-LTR	aara8	3
	Ag-Jock-13	3
	Agam2	4
	Amer3	1
	Baggins	1
	Baggins1	2
	Bilbo	3
	BS	2
	Cr1-1	4
	Cr1-4	2

TE order	TE family	copies (#)
	Copia	16
	Copia1	8
	Copia2/Dm88	7
	Copia3	2
	Copia4	20
	Copia5	2
	diver	8
	diver2	2
	flea	1
	frogger	4
	GATE	1
	Gypsy	2
	Gypsy1	1
	Gypsy10	4
	Gypsy12	1
	Gypsy12A	1
	Gypsy15	1
	Gypsy19	1
	Gypsy20	1
	Gypsy21	5
	Gypsy24	1
	Gypsy25	1
	Gypsy26	2
	Gypsy27	3
	Gypsy29	1
	Gypsy37	2
	Gypsy38	1
	Gypsy4	1
	Gypsy40	2
	Gypsy41	1
	Gypsy5	1

TE order	TE family	copies (#)
	Cr1-8	1
	Cr1-9	2
	Cr1A	1
	Cr1A3	2
	Doc2	2
	Doc3	2
	Doc4	1
	Doc5	1
	Doc6/Juan	4
	Dong	1
	F-element	3
	FW	2
	G2	1
	G3	3
	G4	2
	G5	2
	HidaAg1	1
	I-element	1
	Loa	7
	MinoAg1	1
	R7Ag1	2
	R7AG2	1
	Rt1	3
	Rt1a	1
	Rt2	1
	RtaG4	9
	Tart	20
	Tart-B	1
	uvir	2
	worf	2

Supplementary Table 5: Distance to closest gene from annotated TE.

Distance to closest gene	no gene evidence on contig	0	0-1kb	1-5kb	5-10kb	10-50kb	Total
TE insertions (#)	105	246	55	54	6	2	468

Supplementary Table 6: Core Eukaryotic Genes Mapping Approach (CEGMA) analysis of the *B. antarctica* assembled genome. Group 1 contains the least conserved of the Core Eukaryotic Genes (CEG) and Group 4 the most conserved.

CEG group	Complete proteins (#)	% Complete	Total observed (#)	Average Copy Number	% Orthologs
Complete	**233**	**93.95**	**264**	**1.13**	**9.44**
Group 1	61	92.42	67	1.1	6.56
Group 2	53	94.64	63	1.19	13.21
Group 3	56	91.8	61	1.09	8.93
Group 4	63	96.92	73	1.16	9.52
Partial	**242**	**97.58**	**295**	**1.22**	**15.29**
Group 1	66	100	77	1.17	12.12
Group 2	55	98.21	71	1.29	20
Group 3	58	95.08	70	1.21	15.52
Group 4	63	96.92	77	1.22	14.29

Supplementary Table 7: CEGMA analysis of five genomes.

Species	CEG set	Complete proteins (#)	% Complete	Total observed (#)	Average Copy Number	% Orthologs
Ae. aegypti	Complete	135	54.44	198	1.47	29.63
	Partial	161	64.92	262	1.63	36.65
An. gambiae	Complete	242	97.58	451	1.86	49.17
	Partial	245	98.79	588	2.4	54.29
B. antarctica	Complete	232	93.55	263	1.13	8.62
	Partial	242	97.58	296	1.22	14.46
C. quinquefasciatus	Complete	127	51.21	165	1.3	27.56
	Partial	140	56.45	206	1.47	35.71
D. melanogaster	Complete	241	97.18	279	1.16	13.69
	Partial	245	98.79	293	1.2	15.92

Supplementary Table 8: GC content in the five dipteran species used for comparative analyses

Species	**Genome GC%**	**Coding GC%**
Ae. aegypti	36.2	49.9
An. gambiae	40.9	56.7
B. antarctica	39.0	47.0
C. quinquefasciatus	34.9	55.3
D. melanogaster	40.2	53.4

Supplementary Table 9: Presence of piRNA pathway genes in *B. antarctica* assembly

Dmel name	**Dmel Annotation**	**Dmel symbol**	**function**	**interacting partners***	**ortholog present in *B. antarctica***
Ago3	CG40300	AGO3	piRNA	2 (aub & vret)	yes
Armitage	CG11513	armi	Helicase	4	yes
Aubergine	CG6137	aub	piRNA	15	yes
Krimper	CG15707	krimp	Tudor	?	yes
Piwi	CG6122	piwi	piRNA	5	yes
Rhino	CG10683	rhi	Chromatin	9	no
SpnE	CG3158	spn-E	Helicase	2	yes
Squash	CG4711	squ	Nuclease	1 (aub)	no
Vasa	CG3506	vas	Helicase	4	yes
Zucchini	CG12314	zuc	Nuclease	1 (aub)	no

* Number of interacting partners determined from FlyBase FB2013_06, released November 1st, 2013

Supplementary Table 10: Coding region length summaries for all loci in the annotations from each species.

Species	Minimum	Median	Mean	Maximum	Total bp (inc Ns)
Ae. aegypti	90	1062	1376	33984	22005111
An. gambiae	81	1140	1551	47532	19648704
B. antarctica	69	1008	1403	26004	18619038
C. quinquefasciatus	60	1016	1310	27324	24835191
D. melanogaster	48	1152	1542	68916	20804839

Supplementary Table 11: Coding region length summaries for loci with one-to-one orthologs.

Species	Minimum	Median	Mean	Maximum	Total bp (inc Ns)
Ae. aegypti	159	1473	1847	15903	6618768
An. gambiae	246	1548	2004	16485	7177935
B. antarctica	183	1455	1827	13965	6544527
C. quinquefasciatus	162	1461	1837	16269	6578769
D. melanogaster	246	1623	2038	16683	7300473

Supplementary Table 12: Intron length summaries for all loci in the annotations from each species.

Species	Median	Mean	Maximum
Ae. aegypti	150	3728	329295
An. gambiae	101	1136	196915
B. antarctica	69	333	19461
C. quinquefasciatus	109	1474	88659
D. melanogaster	103	955	142973

Supplementary Methods

Genome size estimate from flow cytometry

Genome size determinations were produced following procedures described in Hare & Johnston [12]; an expansion on those methods is provided here. A single head of the species of interest plus the single head of *D. melanogster* standard (1C = 175 Mbp) were placed into 1 ml of Galbraith buffer in a 2-ml Kontes Dounce homogenizer tube and stroked 15 times with the A pestle to release nuclei from both the sample and standard. The resultant solution was filtered through 40U nylon mesh, stained a minimum of 20 min in the dark with 25 ul of propidium iodide, and then run on a Partec Cyflow cytometer to score relative red fluorescence (> 590 nm) of nuclei from the sample and standard. The amount of DNA in the sample was determined as the mean channel number of the 2C peak of the sample divided by the mean channel number of the 2C peak of the standard times the amount of DNA in the standard. All DNA estimates were determined from a co-preparation of sample and (internal) standard. The position of the sample peak relative to that of the other peaks was established by a single run with the sample or (external) standard prepared and stained individually. Average genome sizes of males and females of *B. antarctica* were based on 20 total replicate estimates on 10 males and 10 females (Supplementary Fig. 1, Supplementary Table 1). Flow cytometry estimates of genome size were also performed for three additional members of the family Chironomidae; the samples were collected in Minnesota and provided by Leonard C. Ferrington Jr., University of Minnesota.

Genome size estimate from sequence reads

Genome size was estimated from sequence reads using a k-mer based approach [13]. The genome size is estimated as the total number of k-mers (in this case 17-mers) divided by the maximal frequency of the k-mer (Supplementary Fig. 2).

De novo genome assembly

Assembly strategy

The assembly individual was sequenced to over 100x coverage using one lane of Illumina HiSeq2000 sequencing technology with a 400 bp insert paired-end sequencing library. A total of 92 million paired-end reads of 101 bp were input into Velvet *de novo* with a k-mer of 55 and an insert length of 400 bp [11]. A total of 5,422 contigs were output from the Velvet *de novo* assembly. Two iterations of ERANGE using the paired-end RNA-sequencing data [14] were used to scaffold the assembled contigs, reducing the number of contigs to 5,064 (Supplementary Fig. 3). Complex regions flanked by sequence reads are represented by stretches of Ns.

Evidence of high quality genome assembly

Multiple lines of evidence confirm that the assembled genome is of high quality and represents most of the DNA sequence. Ninety-five percent of the sequencing reads mapped to the reference genome, with a modal coverage of 177; coverage is calculated based on reads mapped to assembled genome using BamTools [1] (Supplementary Fig. 4). Genome quality assessment was also accomplished by mapping RNA-sequencing data to the assembled genome (see Methods); over 87% of RNA-sequencing reads from Teets et al.[15] mapped to the assembled reference. The RNA-sequencing libraries were exclusively from 4th instar larvae, thus only genes

expressed during the fourth larval instar are present in the data. Moreover, the concordance between the flow cytometry estimate and the assembled genome size suggest that the assembly is complete, with little or no significant blocks of missing chromatin.

Repeats and transposable elements

Transposable element (TE) insertion locations were identified (see Methods, Table 2, Supplementary Data 1). Sixty-eight of the TE insertion sites contain more than one nested TE insertion (Supplementary Table 3). Sequences at the remaining 468 sites clearly correspond to unique TE insertions, representing TEs from the three main TE orders (DNA, non-Long Terminal Repeat (LTR), LTR) (Supplementary Table 4). Most insertions corresponded to retroelements: 306 LTR, 107 non-LTR retroelements, and only 55 DNA elements. No annotated genes were found in contigs containing 105 of the 468 unique TE insertions identified in the assembled genome, indicating that some contigs contain highly repetitive sequence and no apparent coding regions. Among the TE insertions, more than 60% were located inside or less than 1Kb from an identified gene (Supplementary Table 5).

We detected one full length LTR, indicating that while we are able to detect full-length LTRs using the short-read sequence data, LTRs are not present in the other LTR retroelements, thus suggesting that those elements are inactive. Multiple approaches to TE assembly recovered only partial TE sequences, each with high divergence from the canonical consensus TE sequence of the respective families. We conclude that there are very few TEs in the genome, and those that are present are likely to be old and inactive.

No species-specific TEs were detected in the raw reads using ReAs [16].

Gene annotation of core eukaryotic genes

The set of core eukaryotic genes in the assembled genome was identified using CEGMA[17]. Out of the 248 identified as the most highly conserved core eukaryotic genes [17], the assembled *B. antarctica* genome contains 233 (Supplementary Table 6). Including partial matches, the assembled genome contains 97.6% of the highly conserved core eukaryotic genes. Moreover, the average copy number of those genes is low (1.13); both lines of evidence suggest that the assembly is complete and contains the majority of the protein coding sequences, especially compared to other larger insect genomes in which core eukaryotic genes were also identified (Supplementary Table 7). The *D. melanogaster* and *An. gambiae* assembled genomes both contain over 97% of the complete core eukaryotic genes, whereas the *Ae. aegypti* and *C. quinquefasciatus* assembled genomes contain 54.4% and 51.2% of the core genes, respectively, with higher average copy number of 1.47 and 1.3, respectively.

piRNA pathway proteins

The presence of piRNA pathway genes was interrogated using the OrthoMCL [18] comparison between the five species. Three of the loci (*rhino*, *squash*, and *zucchini*) are not present in a cluster of orthologous genes. We executed a tblastx search of the genome for those loci to confirm the absence of the genes in the genome and not solely mis-annotation. The tblastx search revealed few regions of homology that were limited to common protein domains (such as PLD6 for *zuc*), suggesting that the loci are not present in *B. antartica*. It has been shown that *rhino*, *krimper*, and *aubergine* have been subjected to pervasive positive selection in Drosophila species [19]. This suggests that a *rhino* ortholog may be sufficiently divergent to identify orthologs, however, this does not appear to be the case for the other two genes.

Evidence for known infections or symbionts

To determine whether there was any evidence for *Wolbachia* or *Spiroplasma melliferum* in *B. antarctica*, sequence reads were mapped to the *Wolbachia* genome (Genbank NZ_AAQP00000000) [20] and *S. melliferum* genome (Genbank AGBZ01000003) [21]. There was no evidence of either species in *B. antarctica.*

Mitochondrial annotation

Contigs from the original genome assembly were compared using BLAST to mitochondrial genomes of *Drosophila melanogaster* (NC_001709.1) and *Chironomus tepperi* (NC_016167.1) [22] to identify mitochondrial contigs. The reference mitochondrial sequence was present as two contigs in the assembled genome. The two contigs were oriented and merged by homology. Annotation of the mitochondria was accomplished by homology using Mauve [23]. Mitochondrial tRNA genes were identified using the tRNAScan-SE 1.21 server [24] and alignment to other mt-tRNA genes. The annotated mitochondrial sequence is 15,912 bp, with 13 genes and 18 tRNAs.

Supplementary References

1. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-1692 (2011).

2. Allegrucci G, Carchini G, Convey P, Sbordoni V. Evolutionary geographic relationships among orthocladine chironomid midges from maritime Antarctic and sub-Antarctic islands. *Biological Journal of the Linnean Society* **106**, 258-274 (2012).

3. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).

4. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19**, 1390-1394 (2002).

5. Petitpierre E. Molecular cytogenetics and taxonomy of insects, with particular reference to the coleoptera. *International Journal of Insect Morphology and Embryology* **25**, 115-134 (1996).

6. Schmidt-Ott U, Rafiqi AM, Sander K, Johnston JS. Extremely small genomes in two unrelated dipteran insects with shared early developmental traits. *Dev Genes Evol* **219**, 207-210 (2009).

7. Zacharias H. Underreplication of a polytene chromosome arm in the chironomid *Prodiamesa olivacea*. *Chromosoma* **72**, 23-51 (1979).

8. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. *Genome Biol* **14**, R47 (2013).

9. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* **22**, 549-556 (2012).

10. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**, 1117-1123 (2009).

11. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829 (2008).

12. Hare EE, Johnston JS. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol* **772**, 3-12 (2011).

13. Zhang G, *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49-54 (2012).

14. Mortazavi A, *et al.* Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res* **20**, 1740-1747 (2010).

15. Teets NM, *et al.* Gene expression changes governing extreme dehydration tolerance in an Antarctic insect. *Proc Natl Acad Sci U S A* **109**, 20744-20749 (2012).

16. Li R, *et al.* ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. *PLoS Comput Biol* **1**, e43 (2005).

17. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).

18. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178-2189 (2003).

19. Simkin A, Wong A, Poh YP, Theurkauf WE, Jensen JD. Recurrent and recent selective sweeps in the piRNA pathway. *Evolution; international journal of organic evolution* **67**, 1081-1090 (2013).

20. Clark AG, *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**, 203-218 (2007).

21. Alexeev D, *et al.* Application of *Spiroplasma melliferum* proteogenomic profiling for the discovery of virulence factors and pathogenicity mechanisms in host-associated spiroplasmas. *J Proteome Res* **11**, 224-236 (2012).

22. Beckenbach AT. Mitochondrial genome sequences of Nematocera (lower Diptera): evidence of rearrangement following a complete genome duplication in a winter crane fly. *Genome Biol Evol* **4**, 89-101 (2012).

23. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**, 1394-1403 (2004).

24. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).